PerPAS: Topology-Based Single Sample Pathway Analysis Method

Chengyu Liu, Rainer Lehtonen, Sampsa Hautaniemi*

Abstract—Identification of intracellular pathways that play key roles in cancer progression and drug resistance is a prerequisite for developing targeted cancer treatments. The era of personalized medicine calls for computational methods that are able to function with one sample or very small set of samples. Developing such methods is challenging because standard statistical approaches pose several limiting assumptions, such as number of samples, that prevent their application when n approaches to one. We have developed a novel pathway analysis method called PerPAS to estimate activity of pathways at a single sample level by integrating pathway topology information and transcriptomics data. In addition, PerPAS is able to identify altered pathways between cancer and control samples as well as to identify key nodes that contribute to the pathway activity. In our case study using breast cancer data, we show that PerPAS is able to identify highly altered pathways that are associated with patient survival. PerPAS identified four pathways that were associated with patient survival and were successfully validated in at least three independent breast cancer cohorts. In comparison to two other pathway analysis methods that function at a single sample level, PerPAS had superior performance in both synthetic and breast cancer expression datasets. PerPAS is freely available as an R package with full documentation at http://csbi.ltdk.helsinki.fi/pub/czliu/perpas/.

Index Terms—Computational biology, bioinformatics, pathway analysis, integration, gene expression.

1 Introduction

A key finding from large-scale cancer sequencing efforts is that histologically similar cancers may have very different genomic landscapes and treatment responses. Accordingly, cancer genomics data have been increasingly used to identify cancer subtypes and to suggest targeted therapies [1]. For example, in breast cancer, five subtypes have been suggested based on transcriptomics profiling [2]: luminal A, luminal B, HER2-enriched, basal-like and normal breast-like. While these efforts have improved the use of the right treatment for the right patient, samples belonging to these subtypes still have significant heterogeneity at the molecular level. Triple-negative breast cancers, which are the major constituent of the basal-like subtype, have been recently classified into six subtypes with different survival time [3], [4]; luminal A subtype has been categorized into four subtypes [5] and two subtypes have been identified for breast cancer patients with the luminal B signature [6], [7]. Identification of smaller and clinically relevant subtypes calls for computational methods that enable analysis of data from a single or few samples.

Alterations in intracellular pathways can have a drastic effect to efficacy of a therapeutic intervention, in particular, targeted therapies. Thus, a number of pathway analysis methods, such as SPIA [8], DEAP [9], DERA [10] and PATHOME [11], have been developed to pinpoint altered pathways. However, the majority of the existing methods are based on comparison of groups of samples, and their use is limited to settings where the number of samples is sufficiently large to allow statistical inference. Recently, some pathway analysis methods for small sample size have been suggested. PARADIGM uses multi-level data from single or few samples to infer activity of pathways [12]; iPAS quantifies pathway aberration at a single sample level by calculating average distance of a cancer sample from control samples [13]; and Pathifier assigns pathway specific scores that represent deviation from control samples [14]. The main issues with these methods are that they require multi-level data or they consider a pathway as a list of genes and do not take pathway topology into consideration [8], [15], [16].

We have developed a novel computational method called PerPAS (**Per**sonalized **P**athway **A**lteration analysi**S**) for the identification of altered pathways for a single sample based on transcriptomics data. PerPAS uses pathway topology information to quantify contribution of an aberrantly expressed gene to pathways and further to characterize pathway activity. Here, we use both breast cancer and synthetic expression data to demonstrate the performance of PerPAS and to compare it to single-sample based pathway analysis methods iPAS and Pathifier.

2 Results

2.1 Overview of PerPAS

PerPAS is designed to quantify pathway activity at a single sample level. The major steps in the PerPAS approach are: preprocessing transcriptomics and pathway data, quantifying gene contribution to a pathway, and calculating personalized pathway activity scores (Fig. 1). Briefly, PerPAS uses control samples to standardize gene expression profiles and extracts pathways from databases (Fig. 1a). PerPAS quantifies contribution of a gene to a pathway by taking pathway topology, such as bottlenecks, which are defined as nodes with high betweenness centrality [17], and hubs [15],

C. Liu, R. Lehtonen and S. Hautaniemi are with Research Programs Unit, Genome-Scale Biology, Faculty of Medicine, University of Helsinki, Helsinki, Finland
* *S. Hautaniemi is the corresponding author.*

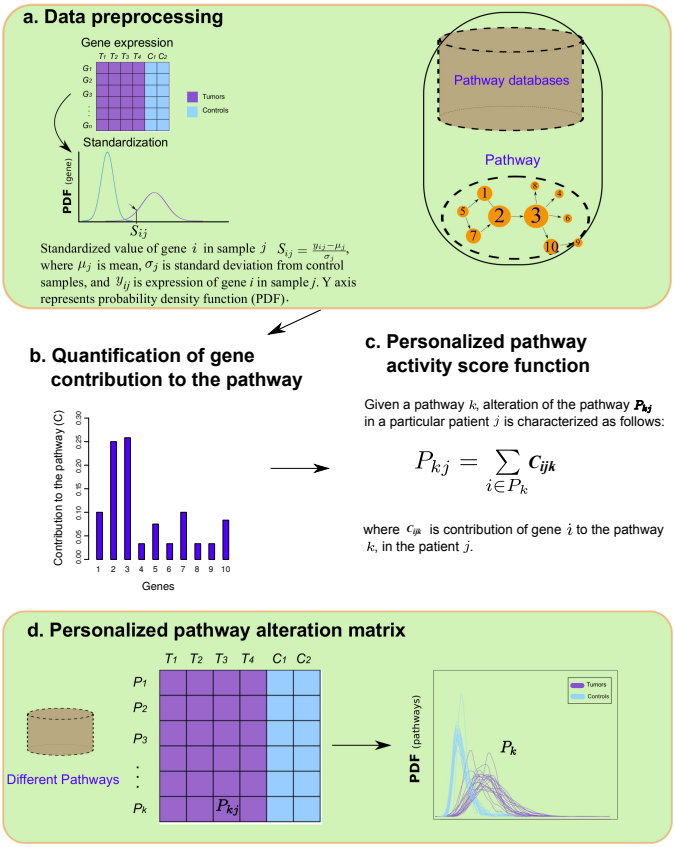

Fig. 1. **Schematic overview of personalized pathway alteration analysis.** Main steps: **a)** Gene expression data are transformed into standardized matrix. From pathway databases, gene regulatory pathways are retrieved. **b)** Gene contribution to a pathway is quantified by taking pathway topology into account. Both topological bottleneck and hub roles are considered. Bottleneck role is quantified by calculating how much signaling the gene mediates while the hub role is modeled by how many downstream genes it directly regulates and how much downstream expression change it induces. **c)** Pathway activity score at an individual sample level is calculated by summing all the gene contribution to the pathway. **d)** Pathway activity score matrix for individual samples is used to identify altered pathways deviated from baseline activity. P_k denotes for a pathway.

[16], into account (Fig. 1b). Bottlenecks measure essentiality of genes in mediating signaling whereas a hub is a highly connected node in a pathway. Quantified gene contribution to a pathway is further summarized to estimate the personalized pathway activity scores (Fig. 1c). The output of PerPAS is a list of pathway activity scores for each sample (Fig. 1d).

2.2 Utility of topology information in pathway analysis

Not all proteins in a pathway are equally important but some, such as bottlenecks [17] and hubs [15], [16], have bigger impact to the pathway activity and outcome than proteins located more peripherally. Accordingly, a central design principle behind PerPAS is to use topological information in the pathway analysis. To demonstrate the importance of topological information and PerPAS, we generated three pathways with varying topology (Fig. 2a) and used PerPAS to indicate the most influential nodes in these pathways.

In the pathway with strong connection (P_c), *Node 2* is a bottleneck (Fig. 2a) and therefore its role in the pathway is stronger than most of the other genes, such as *Node 4*. Obviously, the higher importance of *Node 2* over *Node 4* should be reflected in the results of pathway analysis. PerPAS results indicate that the contribution of *Node 2* in the bottleneck role to the pathway P_c is 8.7 times higher than that of *Node 4* (Fig. 2b). In our study, the contribution of the bottleneck role is represented as a fraction of signaling flows that go through a particular node over all the signaling flows in the pathway. The value ranges from 0 to 1 (See more details in Methods).

Node 3 is another important node since it is a hub and it regulates four downstream genes out of ten genes in the pathway P_c (Fig. 2a). Although *Node 3* itself is not differentially expressed, its downstream genes are (Fig. 2a). Hence, it is valuable to take hub roles into account when performing pathway analysis. PerPAS results indicate the hub role of *Node 3* is not concealed by the fact of its unchanged expression in the pathway P_c (Fig. 2c). In our breast cancer study, the contribution of a hub role is modeled as mean of gene expression of all its direct downstream genes. This allows for identification of nodes with subtle expression changes that do not pass statistical testing but are clearly of biological interest.

By combining the bottleneck and hub roles of genes, PerPAS quantifies the contribution of each gene to the pathway and ranks genes (Fig. 2d). PerPAS results show that *Node 2* and *Node 3* are the most important genes in the pathway P_c, which is consistent with the fact that *Node 2* and *Node 3* are a hub and a bottleneck, respectively.

In addition to ranking nodes in a pathway based on their influence, PerPAS is also able to compare their contribution to different pathways. For instance, in P_b *Node 3* is a hub and activates *Node 8*, whereas in P_c *Node 3* is not only a hub but also a bottleneck. It is the only bridge from *Node 1, 5* and *7* to *Node 4, 6, 8, 9* and *10* (Fig. 2a). The increased importance of *Node 3* to the pathway in P_c is quantified by PerPAS that estimates the contribution of *Node 3* as 0.08 in P_b and 0.20 in P_c (Fig. 2d).

Pathway analysis methods that do not take topology information into account, such as iPAS and Pathifier, will produce identical results for the cases P_a, P_b and P_c. Thus, they are not able to rank *Node 3* higher and thus may fail to identify biologically important nodes.

2.3 Identification of pathways and nexus genes in breast cancer

In order to demonstrate that PerPAS is able to produce robust and potentially important results, we have applied PerPAS to a large breast cancer cohort from The Cancer Genome Atlas (TCGA) ($n = 984$). The most interesting findings were validated in four other breast cancer cohorts GSE1456 ($n = 159$), GSE3494 ($n = 236$), GSE4922 ($n = 249$) and GSE7390 ($n = 198$).

PerPAS identified 40 pathways that were significantly altered in the breast cancer samples compared to the controls in the TCGA cohort (t-test $q < 10^{-60}$). Out of 40 altered pathways, seven were significantly associated with breast cancer patient overall survival (log-rank $p < 0.01$). Four

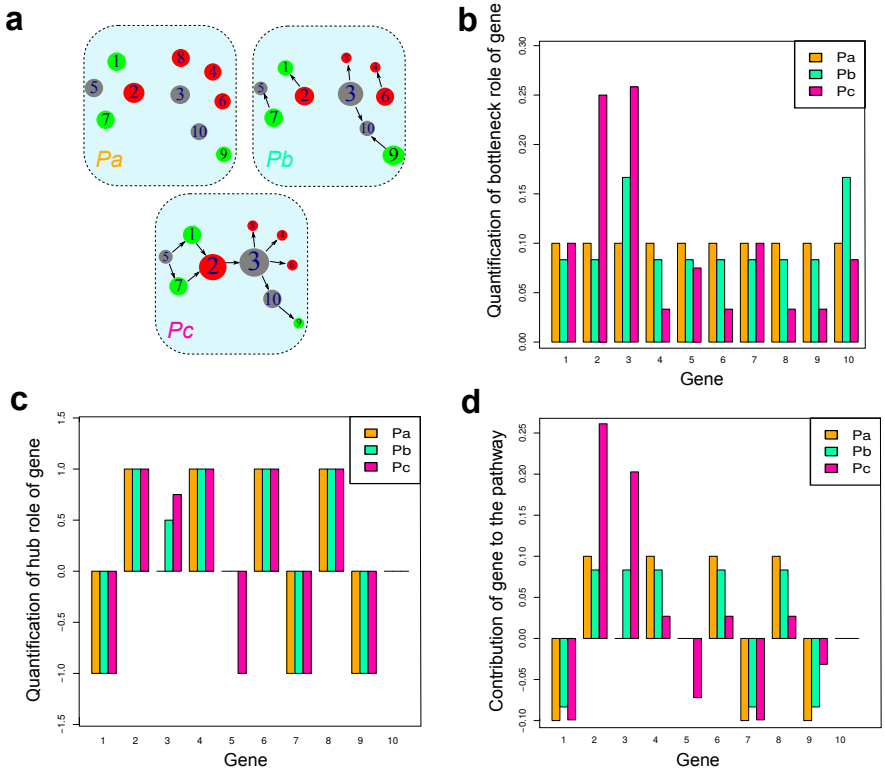

Fig. 2. **Utility of PerPAS on synthetic data. a)** Three synthetic pathways are manually created. Expression values of each gene in three pathways are identical. However, these pathways have different topology, no connection (*Pa*), weak (*Pb*) and strong connection (*Pc*). Node sizes associate with their topological hub role to the pathway. Red, green and gray colored nodes represent over, under and unchanged expression, respectively, compared to control samples. Their corresponding gene expression values are 1, -1, and 0. **b)** Gene contribution to the pathways in bottleneck role. Y axis indicates the percentage of signaling flows going through a particular gene given all signaling flows in the pathway. Contribution of bottleneck role scales from 0 to 1. Zero indicates none of signaling goes through the gene while 1 indicates all the signaling flows go through the gene. **c)** Gene contribution to the pathways in the hub role. Y axis indicates total expression change that the gene induces to its direct downstream genes. Contribution of the hub role can be either negative or positive. Negative values indicate the gene inhibits its downstream genes while positive values indicate the gene activates its downstream genes. **d)** Gene contribution to the pathways. Y axis indicates combined gene contribution of the bottleneck and hub roles.

pathways (Aurora B signaling, growth hormone signaling, *PLK1* signaling events and *LPA4*-mediated signaling events) were validated in at least three independent cohorts (log-rank $p < 0.05$; Supplementary File S1).

Our results show poor survival of patients with high activity of Aurora B signaling or growth hormone signaling pathway. The roles of growth hormone and Aurora B pathways in breast cancer pathogenesis are firmly supported [18], [19], [20].

Interestingly, *PLK1* signaling events pathway showed the most significant survival association and this association was supported in all four independent validation cohorts (Fig.3a-e). Patients with low activity of *PLK1* signaling events pathway have statistically significant survival benefit as compared to patients with high activity of *PLK1* signaling events pathway. The *PLK1* gene plays a critical role in this pathway; out of all 3,450 signaling flows in the *PLK1* signaling events pathway, 979 (28.4%) go through *PLK1* (Supplementary File S2). *PLK1* directly regulates a number of cancer progression driver genes, such as *CDC20/25C* [21], *AURKA* [22], *ECT2* [23], *TPT1* [24] and *BUB1* [25] (Fig. 3f). An example of quantifying gene contribution to *PLK1* signaling events pathway in a single sample is shown in Supplementary File S2.

Our analysis shows a strongly similar expression pattern for *PLK1* downstream genes (Fig. 3f). Furthermore, *PLK1* expression was highly correlated with these direct downstream genes (Fig. 3g). This shows how PerPAS can be used to illustrate pathway activity, which may vary between samples, to identify genes whose expression patterns are tightly co-regulated. For example, many expression values for the majority of genes in the *PLK1* signaling events pathway, e.g., *BUB1*, *CDK1* and *CCNB1*, correlate strongly with *PLK1*. However, there are some genes, such as *WEE1*, *TPT1* and *KIF2A*, whose expression values do not correlate with *PLK1*. For instance, all upstream regulators for *WEE1* (*PLK1*, *CCNB1* and *CDK1*; Fig.3f) correlate with each other whereas *WEE1* itself does not. Regulation of *PLK1* is affected by various mechanisms, such as by phosphorylation of target genes, PLK1 enzyme activity and protein structural variation. Moreover, PLK1 is an essential regulator in many functions which have distinctive regulatory features. For example, in mitotic entry WEE1 is inactivated by PLK1 and further masked by CDC25B, whereas in G2 DNA damage PLK1 is degraded resulting in WEE1 activation [26], [27]. PLK1 and WEE1 inhibitors have shown promising preclinical and clinical effects in targeted and combinatorial therapies in cancers [26], [27]. While further interpretation of any regulatory relationships and mechanisms from any pathway analysis requires additional effort, such as functional exper-

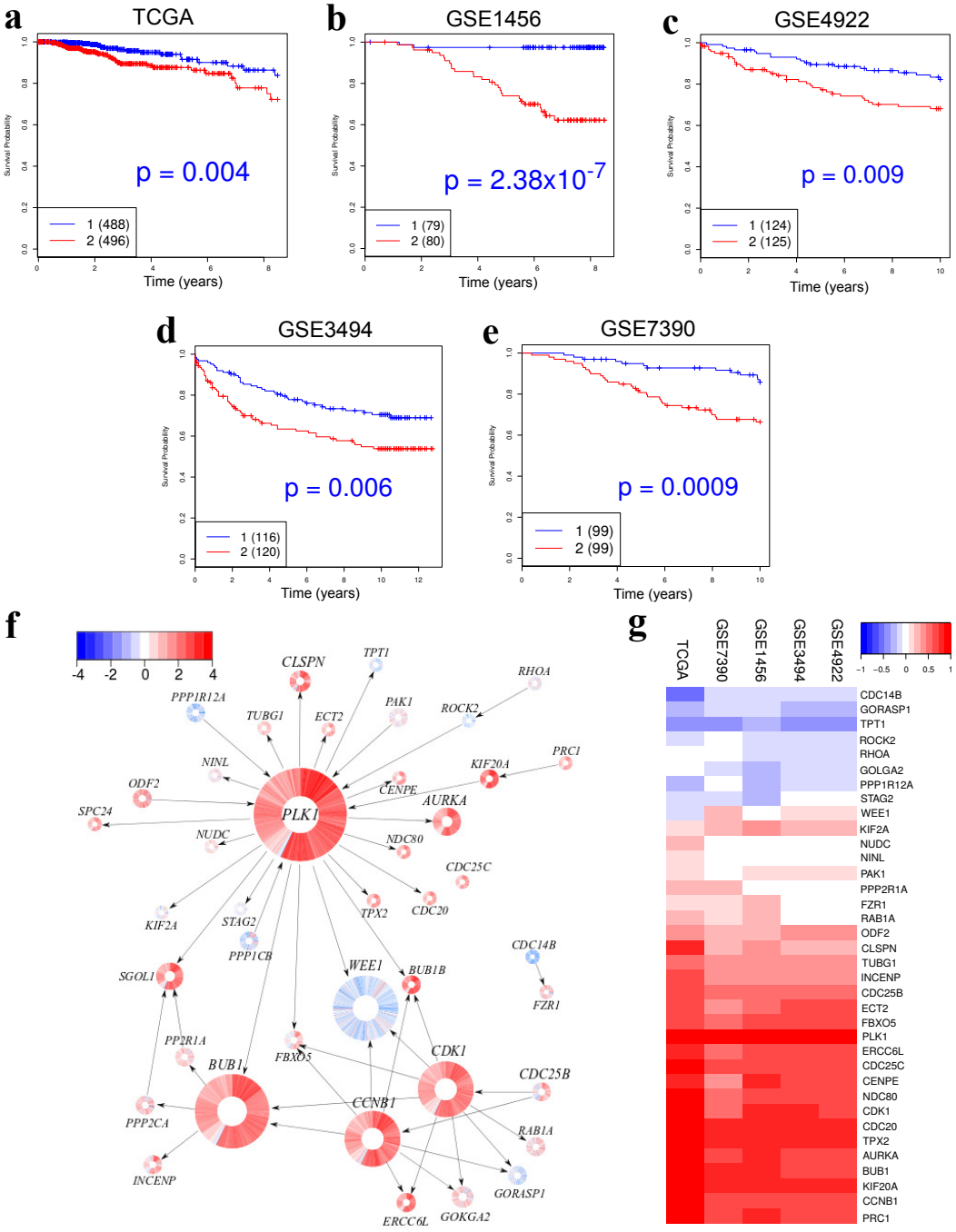

Fig. 3. **Characteristics of PLK1 signaling events pathway. a-e)** Survival association of breast cancer patients to *PLK1* signaling events pathway. *PLK1* signaling events pathway was scored by PerPAS and patients were divided into two groups by median value of personalized pathway activity scores of *PLK1* signaling events pathway. Patients with low- and high-activity of this pathway are represented as 1 (blue) and 2 (red), respectively. Vertical ticks represent censoring events. The X and Y axes represent follow-up time in years and the percentage of survival, respectively. Survival associated p-value was calculated using log-rank test. **f)** *PLK1* signaling events pathway. Each color bar represents a cancer sample and the cancer samples are displayed in the same order for each gene. Color of bar denotes the relative expression of genes in each cancer sample compared to control samples in the TCGA cohort and size of nodes represents the topological importance of the genes. **g)** Pearson correlation between *PLK1* and genes in the pathway in the five cohorts. Note: genes *BUB1B, SGOL1, PPP1CB, PPP2CA* and *SPC24* are excluded in the heatmap since they are not included in the GEO cohorts.

iments, that are out of scope of this study, PerPAS analysis can be used to recognize common regulatory features which are not, at least directly, detectable from RNA expression data.

2.4 Comparison of PerPAS to iPAS and Pathifier using breast cancer data

One way to evaluate the performance of a pathway analysis method is to test whether the identified pathways are associated with patient survival, for example in analyses of prognostic subgroups or drug responses. Significant associ-

TABLE 1
Comparison of survival association between PerPAS, iPAS and Pathifier.

Pathways	TCGA			GSE1456		GSE3494		GSE4922		GSE7390	
	PerPAS	iPAS	Pathifier	PerPAS	iPAS	PerPAS	iPAS	PerPAS	iPAS	PerPAS	iPAS
Growth hormone signaling pathway	0.007	0.001	-	0.00002	0.005	0.005	-	0.004	-	-	-
PLK1 signaling events	0.004	-	-	0.0000007	-	0.009	-	0.006	-	0.0009	-
Aurora B signaling	0.007	-	-	0.00005	-	0.001	-	0.007	-	0.004	-
LPA4-mediated signaling events	0.0005	-	-	0.05	-	0.02	-	0.006	-	-	-
Map kinase signaling,pathway	0.001	-	-	-	-	-	-	-	-	-	-
mechanism of protein import into the nucleus	0.004	-	-	-	-	-	-	-	-	-	-
Signaling Pathways in Glioblastoma	0.01	-	-	-	-	-	-	-	-	-	-
Inhibition of cellular proliferation by gleevec	-	0.007	0.005	-	-	-	0.02	-	-	-	-
Keratinocyte differentiation	-	0.004	-	-	-	-	-	-	-	-	-
Mechanism of gene regulation by peroxisome proliferators via ppara	-	0.0004	-	-	-	-	-	-	-	-	-
Validated targets of C-MYC transcriptional repression	-	0.005	-	-	-	-	-	-	-	-	-
Oncostatin M Signaling Pathway	-	0.0001	-	-	-	-	-	-	-	-	-
Senescence and Autophagy	-	-	0.00005	-	-	-	-	-	-	-	-
HIF-1-alpha transcription factor network	-	-	0.003	-	-	-	-	-	-	-	-
Lissencephaly gene (LIS1) in neuronal migration and development	-	-	0.001	-	-	-	-	-	-	-	-
Reelin signaling pathway	-	-	0.0006	-	-	-	-	-	-	-	-
Regulation of nuclear beta catenin signaling and target gene transcription	-	-	0.008	-	-	-	-	-	-	-	-

Note: Patients were divided into two groups by the median value of the pathway scores calculated by each method. The associated p-value was calculated using log-rank test. "-" denotes p-value that is larger than 0.05.

ation between patient survival and a pathway strongly supports the importance of the pathway in cancer progression or drug resistance. The survival association test, however, is stringent and the lack of survival association does not necessarily mean that a pathway is not important in cancer progression.

To compare the performance of PerPAS to iPAS and Pathifier, we used identical setting, i.e., the TCGA data were used as the discovery cohort and the Gene Expression Omnibus (GEO) data were used as the validation cohorts.

For comparison purpose, we selected similar numbers of significantly altered pathways using different significance thresholds for each method, resulting in 40 significantly altered pathways for PerPAS ($q < 1 \cdot 10^{-60}$), 40 for iPAS ($q < 1 \cdot 10^{-60}$), and 43 pathways for Pathifier ($q < 1 \cdot 10^{-140}$) from the discovery cohort. We used t-test to compare pathway activity scores between cancer and control samples.

iPAS and Pathifier identified six and seven survival associated pathways in the discovery cohort, respectively (Table.1). Four pathways (57%) identified by PerPAS were validate in at least three independent cohorts, whereas from the pathways identified by iPAS two were validated in one independent cohort. One pathway identified by iPAS was also identified by PerPAS (growth hormone signaling pathway). Growth hormone signaling pathway was validated in one validation cohort in the iPAS analysis, whereas PerPAS identified it in three validation cohorts with higher survival association.

Interestingly, the number of common pathways between PerPAS, iPAS and Pathifier is very small. The major reason for the differences between PerPAS and the other two methods is that PerPAS takes topology information into account. This, however, does not explain small number of common pathways between iPAS and Pathifier: only one of the seven pathways identified by Pathifier in the discovery cohort overlapped with the iPAS results (Table 1). While comprehensive comparison between iPAS and Pathifier is out of scope of this study, a reason for discrepancy is that Pathifier uses all samples together to derive a principal curve for each pathway and thus does not give estimates for an individual case-control sample pair like iPAS. Furthermore, pathways identified by Pathifier could not be validated in the independent cohorts because they lacked control samples. The requirement of control samples is a key limitation for Pathifier as not all the cohorts contain control samples.

3 Discussion

Identification of pathways that are altered in tumors compared to controls and drive cancer progression or drug resistance is a prerequisite for personalized medicine. There is a pressing need for pathway analysis methods that work at a single cancer sample level, and are able to pinpoint the most important pathways and their central nodes in an individual samples. However, most of existing pathway analysis methods compare two or more groups of samples and do not support pathway analysis at a single sample level. Further, those pathway analysis methods that support analysis of a single sample do not integrate pathway topology. Herein presented PerPAS allows both single-sample analysis and takes network topology into account.

We have used both synthetic and breast cancer expression data to demonstrate the utility of PerPAS. Results from synthetic data demonstrated that PerPAS is able to prioritize nodes that are central for the network signaling. In the breast cancer data, PerPAS identified seven pathways with survival association in the discovery cohort from which four were validated in at least three independent validation cohorts. While pathway's survival association is a stringent criterion, it is one of the most useful tests for pathway methods as the users of pathway methods typically are interested in finding pathways that may have clinical significance. *PLK1* signaling events pathway was associated with survival in all five breast cancer cohorts. PerPAS highlighted the *PLK1* gene as a central node in the pathway, it was also highly correlated with most of its downstream genes.

Standardization of gene expression in PerPAS is an important step to minimize cohort effects. It provides comparability between expression data and between results from the data. In an ideal case, a set of control samples from the same cohort is used to standardize gene expression of treatment samples. PerPAS can also be used without control samples, for example, by standardizing gene expression to the mean of the cohort or by skipping standardization step

in case only one or a few samples are available. PerPAS is applicable to conduct pathway analysis for gene expression data from any disease. We have shown here that PerPAS is applicable to gene expression data. However, PerPAS should also be applicable to many other molecular data, such as protein-protein interaction data. PerPAS requires molecular measurements and networks as its inputs.

In summary, we have developed a novel pathway analysis method, PerPAS, that is optimized for single sample analysis. PerPAS uses pathway topology information to quantify pathway activity scores, and to identify aberrant pathways and key nodes in the pathways. Our results show that survival associated pathways identified by PerPAS have a much higher rate of being validated in the independent validation cohorts than the other single-sample level pathway methods.

4 METHODS

4.1 PerPAS

4.1.1 Data preprocessing

Data preprocessing consists of two steps: standardization of gene expression and preparation of pathways.

4.1.1.1 Standardization of gene expression: We adopted the method used by Maxime *et al.* [13] to standardize gene expression. Briefly, for expression cohorts that contain control samples (e.g., TCGA data), gene expression of a tumor sample is standardized to the mean and the standard deviation of control samples. We extended this method to gene expression cohorts that lack control samples (e.g., GEO data sets). Gene expression is normalized by the mean and the standard deviation of the cohort, instead of using external control samples. This measures gene expression difference between a tumor sample and the mean of gene expression of all the tumor samples in the cohort.

4.1.1.2 Preparation of pathways: We obtained pathways from NCI-Nature Pathway Interaction Database (PID) [28] and WikiPathways [29]. Level 3 biopax-formatted PID was analyzed using rBiopaxParser package [30]. We obtained WikiPathways from Moksiskaan database where many useful application programming interfaces (APIs) are provided to ease extraction of different types of interactions, such as gene activation and inhibition interactions for each pathway [31]. We excluded pathways with less than four nodes resulting in 368 pathways from the PID and 75 pathways from WikiPathways.

4.1.2 Quantification of gene contribution to a pathway

Some nodes in a pathway are more central and important than others. Examples of such central nodes are hubs [15], [16] and bottlenecks [17]. Contribution of a gene to a pathway can be quantified according to its hub and bottleneck roles. Impact of a hubness of a gene can be evaluated by measuring the number of genes it directly regulates and expression changes of its direct downstream genes. Quantification of the hub role of a gene in a pathway *k* is shown in Eq. 1,

$$H_{ijk} = \frac{\sum_{l \in \{\text{direct downstream genes of gene } i \text{ in the pathway } k\}} S_{lj}}{M_i} \quad (1)$$

where S_{lj} is standardized expression of gene l in sample j and M_i is the number of direct downstream genes of gene i. In cases where there are no more than two direct downstream genes for a gene, a hub score is represented by its own standardized expression.

Bottleneck measures essentiality of genes in controlling signaling flows in the pathway [17]. Bottleneck role of a gene can be quantified by estimating the percentage of signaling flows that go through the gene over all the signaling flows. A signaling flow can be considered as a path from one gene to another. Identification of all possible paths from one gene to another in a pathway is computationally costly and furthermore identifying all the paths between all genes is infeasible, especially when the pathway is complicated. Hence, we use a shortest path to represent a signaling flow between two genes [17], which leads to an assumption that is signaling from one gene to another always transmits through the shortest path. The shortest path is a path between two nodes where the sum of weights of its constituent edges is minimal. In our method, weights of edges are equal and thus, can be ignored. A breadth-first search algorithm is used to find shortest paths between any two genes in a pathway. Quantification of bottleneck role of a gene is shown in Eq. 2,

$$Q_{ik} = \frac{n_{ik}}{N_k}, \quad (2)$$

where N_k is the total number of signaling flows in pathway k and n_{ik} is the number of signaling flows to which gene i contributes. Value of Q_{ik} ranges from zero to one.

Finally, given a case sample j, topological contribution of gene i to pathway k is represented by multiplication of hub and bottleneck roles of the gene shown in Eq. 3.

$$C_{ijk} = Q_{ik} \cdot H_{ijk} \quad (3)$$

4.1.3 Personalized pathway activity score

Personalized pathway activity is the activity of a pathway in a particular sample. It is summarized from the topological contribution of all genes in the pathway. Given a pathway k and a sample j, personalized pathway activity score P_{kj} is defined as follows:

$$P_{kj} = \sum_{i \in pathway_k} C_{ijk} \quad (4)$$

To assess the significance of personalized pathway activity score, two permutation tests are performed under the null hypothesis of "personalized pathway activity score is random". In the first test, PerPAS is applied over 100 random trials where the gene expression of the pathway in the sample is randomly permuted. The second test randomly assigns gene regulations (edges) to any two genes in the pathway followed by PerPAS scoring. This procedure is repeated 100 times. The significant level of both tests is calculated by comparing the observed score to the mean of random scores on permutations. Both ways of randomization disrupt expression correlation between genes of a gene regulation. In the first permutation, the topology of pathway remains, and hence the test answers how randomizing expression of genes changes personalized pathway activity score. The second test permutes pathway topology and

thus tests how topology influences personalized pathway activity score.

4.2 Breast cancer data

Log2 transformed level 3 RNA-seq gene expression data were downloaded from The Cancer Genome Atlas (TCGA) repository [32] and were used as the discovery cohort. We discarded samples without survival time or vital status information, resulting in 984 breast cancer samples and 111 control samples. For validation, we used four publicly available breast cancer data cohorts from Gene Expression Omnibus (GEO) [33]: GSE1456 ($n = 159$), GSE3494 ($n = 251$), GSE4922 ($n = 236$) and GSE7390 ($n = 198$). For these GEO data cohorts, gene level normalization was performed by using Robust Multi-array Average (RMA) and the data were log2 transformed.

ACKNOWLEDGMENT

The results published here are in part based upon data generated by The Cancer Genome Atlas pilot project established by the NCI and NHGRI. Information about TCGA and the investigators and institutions who constitute the TCGA research network can be found at http://cancergenome.nih.gov . We are grateful to CSC — IT Center for Science Ltd. for compute resources.

This work was supported financially by Academy of Finland (Center of Excellence in Cancer Genetics Research), Sigrid Jusélius foundation, Finnish Cancer Associations and Integrative Life Science Graduate Program (ILS).

REFERENCES

[1] S. J. Schnitt, "Classification and prognosis of invasive breast cancer: from morphology to molecular taxonomy," *Modern Pathology: An Official Journal of the United States and Canadian Academy of Pathology, Inc*, vol. 23 Suppl 2, pp. S60–64, May 2010.

[2] T. Sorlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geysers, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lonning, and A. L. Borresen-Dale, "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 19, pp. 10 869–10 874, Sep. 2001.

[3] B. D. Lehmann, J. A. Bauer, X. Chen, M. E. Sanders, A. B. Chakravarthy, Y. Shyr, and J. A. Pietenpol, "Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies," *The Journal of Clinical Investigation*, vol. 121, no. 7, pp. 2750–2767, Jul. 2011.

[4] R. T. Lawrence, E. M. Perez, D. Hernandez, C. P. Miller, K. M. Haas, H. Y. Irie, S.-I. Lee, C. A. Blau, and J. Villen, "The proteomic landscape of triple-negative breast cancer," *Cell Reports*, vol. 11, no. 4, pp. 630–644, Apr. 2015.

[5] G. Ciriello, R. Sinha, K. A. Hoadley, A. S. Jacobsen, B. Reva, C. M. Perou, C. Sander, and N. Schultz, "The molecular diversity of Luminal A breast tumors," *Breast Cancer Research and Treatment*, vol. 141, no. 3, pp. 409–420, Oct. 2013.

[6] C. J. Creighton, "The molecular profile of luminal B breast cancer," *Biologics: Targets & Therapy*, vol. 6, pp. 289–297, 2012.

[7] F. Ades, D. Zardavas, I. Bozovic-Spasojevic, L. Pugliano, D. Fumagalli, E. de Azambuja, G. Viale, C. Sotiriou, and M. Piccart, "Luminal B breast cancer: molecular characterization, clinical management, and future perspectives," *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, vol. 32, no. 25, pp. 2794–2803, Sep. 2014.

[8] A. L. Tarca, S. Draghici, P. Khatri, S. S. Hassan, P. Mittal, J.-S. Kim, C. J. Kim, J. P. Kusanovic, and R. Romero, "A novel signaling pathway impact analysis," *Bioinformatics (Oxford, England)*, vol. 25, no. 1, pp. 75–82, Jan. 2009.

[9] W. A. Haynes, R. Higdon, L. Stanberry, D. Collins, and E. Kolker, "Differential Expression Analysis for Pathways," *PLoS Comput Biol*, vol. 9, no. 3, p. e1002967, Mar. 2013.

[10] C. Liu, R. Louhimo, M. Laakso, R. Lehtonen, and S. Hautaniemi, "Identification of sample-specific regulations using integrative network level analysis," *BMC cancer*, vol. 15, p. 319, 2015.

[11] S. Nam, H. R. Chang, K.-T. Kim, M.-C. Kook, D. Hong, C. H. Kwon, H. R. Jung, H. S. Park, G. Powis, H. Liang, T. Park, and Y. H. Kim, "PATHOME: an algorithm for accurately detecting differentially expressed subpathways," *Oncogene*, vol. 33, no. 41, pp. 4941–4951, Oct. 2014.

[12] C. J. Vaske, S. C. Benz, J. Z. Sanborn, D. Earl, C. Szeto, J. Zhu, D. Haussler, and J. M. Stuart, "Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM," *Bioinformatics (Oxford, England)*, vol. 26, no. 12, pp. i237–245, Jun. 2010.

[13] T. Ahn, E. Lee, N. Huh, and T. Park, "Personalized identification of altered pathways in cancer using accumulated normal tissue data," *Bioinformatics (Oxford, England)*, vol. 30, no. 17, pp. i422–429, Sep. 2014.

[14] Y. Drier, M. Sheffer, and E. Domany, "Pathway-based personalized analysis of cancer," *Proceedings of the National Academy of Sciences*, vol. 110, no. 16, pp. 6388–6393, Apr. 2013.

[15] H. Jeong, S. P. Mason, A.-L. Barab??si, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, pp. 41–42, May 2001.

[16] R. Albert, H. Jeong, and A.-L. Barabasi, "Error and attack tolerance of complex networks : Article : Nature," *Nature*, vol. 406, no. 6794, pp. 378–382, Jul. 2000.

[17] H. Yu, P. M. Kim, E. Sprecher, V. Trifonov, and M. Gerstein, "The Importance of Bottlenecks in Protein Networks: Correlation with Gene Essentiality and Expression Dynamics," *PLOS Comput Biol*, vol. 3, no. 4, p. e59, Apr. 2007.

[18] D. S. Boss, J. H. Beijnen, and J. H. M. Schellens, "Clinical experience with aurora kinase inhibitors: a review," *The Oncologist*, vol. 14, no. 8, pp. 780–793, Aug. 2009.

[19] J. F. Hilton and G. I. Shapiro, "Aurora kinase inhibition as an anticancer strategy," *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, vol. 32, no. 1, pp. 57–59, Jan. 2014.

[20] R. Dent, M. Trudeau, K. I. Pritchard, W. M. Hanna, H. K. Kahn, C. A. Sawka, L. A. Lickley, E. Rawlinson, P. Sun, and S. A. Narod, "Triple-Negative Breast Cancer: Clinical Features and Patterns of Recurrence," *Clinical Cancer Research*, vol. 13, no. 15, pp. 4429–4434, Aug. 2007.

[21] R. Bagheri-Yarmand, A. Nanos-Webb, A. Biernacka, T. Bui, and K. Keyomarsi, "Cyclin E deregulation impairs mitotic progression through premature activation of Cdc25c," *Cancer Research*, vol. 70, no. 12, pp. 5085–5095, Jun. 2010.

[22] N. Ertych, A. Stolz, A. Stenzinger, W. Weichert, S. Kaulfu??, P. Burfeind, A. Aigner, L. Wordeman, and H. Bastians, "Increased microtubule assembly rates influence chromosomal instability in colorectal cancer cells," *Nature Cell Biology*, vol. 16, no. 8, pp. 779–791, Aug. 2014.

[23] V. Justilien, L. Jameison, C. J. Der, K. L. Rossman, and A. P. Fields, "Oncogenic activity of Ect2 is regulated through protein kinase C iota-mediated phosphorylation," *The Journal of Biological Chemistry*, vol. 286, no. 10, pp. 8149–8157, Mar. 2011.

[24] R. Amson, S. Pece, J.-C. Marine, P. P. Di Fiore, and A. Telerman, "TPT1/ TCTP-regulated pathways in phenotypic reprogramming," *Trends in Cell Biology*, vol. 23, no. 1, pp. 37–46, Jan. 2013.

[25] Y. Ding, C. G. Hubert, J. Herman, P. Corrin, C. M. Toledo, K. Skutt-Kakaria, J. Vazquez, R. Basom, B. Zhang, J. K. Risler, S. M. Pollard, D.-H. Nam, J. J. Delrow, J. Zhu, J. Lee, J. DeLuca, J. M. Olson, and P. J. Paddison, "Cancer-Specific requirement for BUB1b/BUBR1 in human brain tumor isolates and genetically transformed cells," *Cancer Discovery*, vol. 3, no. 2, pp. 198–211, Feb. 2013.

[26] B. T. Gjertsen and P. Sch??ffski, "Discovery and development of the Polo-like kinase inhibitor volasertib in cancer therapy," *Leukemia*, vol. 29, no. 1, pp. 11–19, Jan. 2015, 00040.

[27] P. C. D. W. Hamer, S. E. Mir, D. Noske, C. J. F. V. Noorden, and T. W??rdinger, "WEE1 Kinase Targeting Combined with DNA-Damaging Cancer Therapy Catalyzes Mitotic Catastrophe," *American Association for Cancer Research*, vol. 17, no. 13, pp. 4200–4207, Jul. 2011, 00088.

[28] C. F. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, and K. H. Buetow, "PID: the Pathway Interaction Database," *Nucleic Acids Research*, vol. 37, no. Database issue, pp. D674–679, Jan. 2009.

[29] T. Kelder, M. P. van Iersel, K. Hanspers, M. Kutmon, B. R. Conklin, C. T. Evelo, and A. R. Pico, "WikiPathways: building research communities on biological pathways," *Nucleic Acids Research*, vol. 40, no. Database issue, pp. D1301–D1307, Jan. 2012.

[30] F. Kramer, M. Bayerlov??, F. Klemm, A. Bleckmann, and T. Beissbarth, "rBiopaxParser–an R package to parse, modify and visualize BioPAX data," *Bioinformatics (Oxford, England)*, vol. 29, no. 4, pp. 520–522, Feb. 2013.

[31] M. Laakso and S. Hautaniemi, "Integrative platform to translate gene sets to networks," *Bioinformatics (Oxford, England)*, vol. 26, no. 14, pp. 1802–1803, Jul. 2010.

[32] Cancer Genome Atlas Network, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, pp. 61–70, Oct. 2012.

[33] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Research*, vol. 30, no. 1, pp. 207–210, Jan. 2002.